445 Cheatsheet

Below is a quick refresher on some math tools and problem-solving techniques from 240 (or other prereqs) that we'll assume knowledge of for the PSets. Roadmap:

- Section 1 is a refresher on probability. I hope that most of it feels like a review. Section 1.8 may feel conceptually more advanced than the rest. We'll use a lot of probability in this course, so you may want to have this handy as a reference if you feel stuck on a problem because of probability.
- Section 2 is a refresher on continuous optimization. You don't need much for this course, but you may find it as a helpful quick reference.
- Section 4 contains my own (brief) thoughts on how to write solid proofs. This may be different than thoughts you've heard in previous courses. **I strongly recommend reading this section**, and in particular the examples demonstrating common pitfalls.
- Section 3 contains quick tips on how to approach creative problem-solving. **I strongly recommend reading this section, especially Section 3.1**, which describes how I think of partial progress/credit.

All of the notes here are intended for an audience who has seen these topics before, but may enjoy a refresher before using these tools to solve abstract problems. If you have not seen a topic before and want to learn it, I suggest revisiting the (significantly more thorough) course materials from 240.

1 Basic Probability

1.1 Discrete random variables

A *random variable* is a variable whose value is uncertain (i.e. the roll of a die). If X is a random variable that always takes non-negative, integer values, (we'll refer to this as a discrete random variable) then we can write the expected value of X as:

$$\textbf{Definition of expected value, form 1: } \mathbb{E}[X] = \sum_{i=0}^{\infty} \Pr[X = i] \cdot i.$$

Probably the above definition is familiar to most of you already. Another way to compute the expected value (which sometimes results in simpler calculations) is:

$$\textbf{Definition of expected value, form 2: } \mathbb{E}[X] = \sum_{i=0}^{\infty} \Pr[X > i].$$

Let's quickly see why the two definitions are equivalent:

$$\sum_{i=0}^{\infty} \Pr[X > i] = \sum_{i=0}^{\infty} \sum_{j>i} \Pr[X = j].$$

$$= \sum_{j=0}^{\infty} \sum_{i<j} \Pr[X = j]$$

$$= \sum_{j=0}^{\infty} j \cdot \Pr[X = j].$$

We obtain the second equality just by flipping the order of sums: the term $\Pr[X = j]$ is summed once for every $i < j$. The third equality is obtained by just observing that there are exactly j non-negative integers less than j.

Why bother knowing two definitions? Most students find it conceptually easier to remember form 1. I also find this to be the case. However, many problems you'll encounter in this course will be conceptually easier to solve using form 2. See the later sections for further discussion and a few examples.

1.2 Continuous random variables

Now let's consider a continuous, non-negative random variable with probability density function (PDF) $f(\cdot)$ and cumulative distribution function (CDF) $F(\cdot)$. What do all these words mean? You should imagine the following mapping:

- **Continuous** just means that the random variable might take any non-negative value. For instance, rather than the roll of a die, a random variable might be the number of seconds you spend reading this sentence.
- The **PDF** is just a formal way of discussing the probability that $X = x$. Because the random variable is continuous, the probability that $X = x$ is actually zero for all x (what is the probability that you spend exactly 3.4284203 seconds reading this sentence)? So we think of dx as being infinitesimally small (the same dx from your calculus classes), and think of $\Pr[X = x]$ as $f(x)dx$. Or, to be even more formal, this captures $f(x) = \lim_{\varepsilon \to 0} \Pr[X \in [x, x + \varepsilon]]/\varepsilon$.
- The **CDF** of a random variable is simpler to define, and just denotes $F(x) = \Pr[X \leq x]$. Note that we therefore have $F(x) = \int_0^x f(y)dy$. Think of this as "summing" (integrating) over all $y \leq x$ the probability that $X = y$ ($f(y)dy$). Therefore $F'(x) = f(x)$ (by fundamental theorem of calculus).

So how do we take the expectation of a continuous random variable? We just need to map the definitions above into the new language.

Definition of expected value, continuous random variables, form 1: $\mathbb{E}[X] = \int_0^{\infty} xf(x)dx.$

You should parse exactly the same way as form 1 for discrete random variables, except we've replaced the sum with an integral, and $\Pr[X = i]$ is now "$f(x)dx \approx \Pr[X = x]$." The equivalent definition for form 2 is also often easier to use in calculations:

Definition of expected value, continuous random variables, form 2: $\mathbb{E}[X] = \int_0^{\infty} (1-F(x))dx.$

If $F(x) = \Pr[X \leq x]$, then $1 - F(x) = \Pr[X > x]$, so this is the same as form 2 for discrete random variables, except we've replaced the sum with an integral. For form 2, it is *crucial* that the integral start below at 0, even when the random variable only takes values (say) > 1. We'll see this in examples below.

1.2.1 Brief Discussion

Note also that the definition of expected value for continuous random variables, form 2, is actually well-defined for any non-negative random variable, whether it is discrete or continuous (or a mix of both). This is not always the mathematically cleanest way to compute the expectation of a complex random variable, but it is the conceptually most-straight-forward. **A safe approach for any random variable X you see in this class is to explicitly figure out, for any x, what is the probability that $X > x$? Then, write this integral and compute it.**

If you're conceptually more comfortable with probability, you may often be able to find a more clever approach which requires fewer calculations, but the above approach will always succeed.

Finally, also note that for continuous random variables, $\Pr[X = x] = 0$ for all x, so $\Pr[X \leq x] = \Pr[X < x]$. This is not true for discrete random variables, but observe that the evaluation of the integral is indifferent to whether we use $F(x) = \Pr[X \leq x]$ or $F(x) = \Pr[X < x]$.[1]

1.3 Three Examples

Example 1 *Let X be a random variable that is 4 with probability $1/2$, and 5 with probability $1/2$ (this is the uniform distribution over $\{4, 5\}$).*

Fact 1 *The expected value of X is 4.5 ($\mathbb{E}[X] = 4.5$).*

We provide two proofs of this, one using each form.

Proof. The expected value as computed by form 1 is:

$$\sum_{i=0}^{\infty} \Pr[X = i] \cdot i = 4 \cdot 1/2 + 5 \cdot 1/2 = 4.5.$$

■

Proof. The expected value as computed by form 2 is:

$$\sum_{i=0}^{\infty} \Pr[X > i] = \sum_{i=0}^{3} 1 + \sum_{i=4}^{4} 1/2 = 4.5.$$

■

Example 2 *Let X be a random variable drawn from the uniform distribution on the interval $[4, 5]$. That is, the PDF of X satisfies $f(x) = 1$ for all $x \in [4, 5]$, and $f(x) = 0$ otherwise. Observe that the CDF of X satisfies $F(x) = 0$ for all $x \in [0, 4]$, $F(x) = x - 4$ for all $x \in [4, 5]$, and $F(x) = 1$ for all $x \geq 5$.*

Fact 2 $\mathbb{E}[X] = 4.5$.

[1] This is because the two definitions differ on a set of measure zero, so the integrals must be identical. But don't worry about this if you're not familiar with this language.

We again provide two proofs, one using each form.

Proof. We can compute the expected value by form 1 as:

$$\int_0^\infty xf(x)dx = \int_4^5 xdx = x^2/2|_4^5 = 25/2 - 8 = 4.5.$$

■

Proof. We can also compute it using form 2 as:

$$\int_0^\infty (1 - F(x))dx = \int_0^4 1dx + \int_4^5 (5 - x)dx + \int_5^\infty 0dx = 4 + ((5x - x^2/2)|_4^5) + 0 = 4.5.$$

■

Note that it is *crucial* that we started the integral at 0 and not 4 for form 2, otherwise we would have incorrectly computed the expectation as $.5$ instead of 4.5. This isn't crucial for form 1, since all the terms in $[0, 4]$ drop out anyway as $f(x) = 0$.

Example 3 *Let X be defined as follows. A fair six-sided die is rolled. Let Y denote the roll, then $X = Y^2$ (that is, if the roll is 1, $X = 1$. If the roll is 2, then $X = 4$, and so on).*

Fact 3 $\mathbb{E}[X] = 91/6$.

Proof. This is an example of a random variable where it happens to be significantly cleaner to compute the expectation using form 1. Here we can write:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{\infty} i \cdot \Pr[X = i] \\
&= \sum_{i=1}^{\infty} i \cdot \Pr[Y = \sqrt{i}] && \text{(because } X = Y^2\text{)} \\
&= \sum_{j=1}^{6} j^2 \cdot \Pr[Y = j] && \text{(substituting } j = \sqrt{i}\text{, and observing that } \Pr[Y = j] = 0 \text{ when } j \notin [6]\text{)} \\
&= \sum_{j=1}^{6} \frac{1}{6} \cdot j^2 = 91/6
\end{aligned}$$

■

1.4 Linearity of Expectation

Linearity of expectation refers to the following simple, but surprisingly useful fact. Let X_1 and X_2 be two random variables. Then $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$. The proof is immediate from the definitions above. We include the proof for the discrete case:

$$
\begin{aligned}
\mathbb{E}[X_1 + X_2] &= \sum_{i=0}^{\infty} \Pr[X_1 + X_2 = i] \cdot i \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{i} \Pr[X_1 = j] \cdot \Pr[X_2 = i - j] \cdot i \\
&= \sum_{j=0}^{\infty} \sum_{i=j}^{\infty} \Pr[X_1 = j] \cdot \Pr[X_2 = i - j] \cdot i \\
&= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot \sum_{\ell=0}^{\infty} \Pr[X_2 = \ell] \cdot (\ell + j) \qquad \text{(changing variables with } \ell = i - j) \\
&= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot \left(j + \sum_{\ell=0}^{\infty} \Pr[X_2 = \ell] \cdot \ell \right) \\
&= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot (j + \mathbb{E}[X_2]) \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2]. \qquad \text{(because } \sum_{j=0}^{\infty} \Pr[X_1 = j] = 1)
\end{aligned}
$$

1.5 Bayes' Rule

Let's first recap the definition of conditional probability: the probability of an event X conditioned on another event Y, denoted by $\Pr[X|Y]$ is equal to the probability of X and Y divided by the probability of Y (that is, $\Pr[X \wedge Y] / \Pr[Y]$). Think of this as the probability that X occurs, given that Y has occured. For a concrete example, consider the probability that a fair six-sided dice lands two (X), conditioned on it landing even (Y). Then $\Pr[X \wedge Y]$ is the probability that the die lands two *and* that it is even (which is just the probability that it is two), so $1/6$. $\Pr[Y]$ is just the probability that the die is even, which is $1/2$, so the ratio is $1/3$. The probability that the roll is prime (X), conditioned on being even (Y) can be computed similarly: the probability that the roll is prime *and* even is $1/6$ (the only even prime is two), and the probability that the roll is even is $1/2$. So again the ratio is $1/3$, and the probability of rolling prime conditioned on rolling even is $1/3$.

Sometimes, explicitly computing $\Pr[X \wedge Y]$ might be challenging, but computing $\Pr[Y|X]$ is not so bad. Bayes' rule simply manipulates the above equalities to write:

$$\Pr[X|Y] = \Pr[X \wedge Y] / \Pr[Y] = \Pr[Y|X] \cdot \Pr[X] / \Pr[Y].$$

Example 4 *Consider a coin whose probability of outputting heads is p, but p is a random variable. Specifically, p is either equal to $1/4$ with probability $1/2$, and equal to $3/4$ with probability $1/2$. You flip the coin once and it lands heads. What is the probability that the coin's bias is $3/4$?*

Fact 4 *The probability that the coin's bias is $3/4$ is $3/4$.*

Proof. We prove this using Bayes' rule. Let X denote the event that $p = 3/4$, and let Y denote the event that the coin lands hands after one flip. Then the problem is asking us to compute

$\Pr[X|Y]$. This is conceptually quite tricky to reason about! But fortunately, Bayes' rule gives us a formula using three terms that are much simpler to think about.

- $\Pr[Y|X]$ is the probability that the coin lands heads, conditioned on the bias being $3/4$. This is easy to compute, and is just $3/4$, immediately from the definition of bias.
- $\Pr[X]$ is the probability that the bias is $3/4$, which is just $1/2$, immediately by the problem setup.
- $\Pr[Y]$ is the probability that the coin lands heads (without having seen any flips), which is also just $1/2$. This is because the coin's probability of landing on heads is always p, and $\mathbb{E}[p] = 1/2$.

Now, we can just apply Bayes' rule and write:

$$\Pr[X|Y] = \frac{\Pr[Y|X] \cdot \Pr[Y]}{\Pr[X]} = \frac{(3/4) \cdot (1/2)}{(1/2)} = 3/4.$$

■

1.6 Conditional Expectation

The previous section covers conditional probability, what is the probability that of X conditioned on Y. We may also want to discuss conditional expectations: the expectation of a random variable X conditioned on an event Y. Formally, think of this as first, "what is the probability that $X = x$, conditioned on Y? Then define a new random variable $X|Y$, which is equal to x with this probability, and take it's expectation.

This is related to the following notion: let $\mathbb{I}(Y)$ denote a random variable which is one when event Y occurs, and 0 otherwise. **Then $\mathbb{E}[X|Y] = \mathbb{E}[X \cdot \mathbb{I}(Y)]/\Pr[Y]$. Sometimes, it may be mathematically simpler to reason about $\mathbb{E}[X \cdot \mathbb{I}(Y)]$, and then $\Pr[Y]$ separately, to compute $\mathbb{E}[X|Y]$.**

1.6.1 An Example

Example 5 *Let X be a random variable that is drawn from a distribution with CDF $1 - e^{-x}$ (this is called an exponential distribution). Observe that the PDF of this distribution is exactly e^{-x} (by taking the derivative of $1 - e^{-x}$). Compute $\mathbb{E}[X|X \in [5, 10]]$.*

Fact 5 $\mathbb{E}[X|X \in [5, 10]] = 6 - 5 \cdot \frac{1}{e^5 - 1}$.

Proof. We propose doing this computation by first computing $\mathbb{E}[X \cdot \mathbb{I}(X \in [5, 10])]$, and then dividing by $\Pr[X \in [5, 10]]$.

Following the bold-faced suggestion in Section 1.2.1, we have a random variable whose expectation we wish to compute. So, let's do so by trying to directly compute $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x]$, for all x.

Observation 1 *For all $x < 5$, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = e^{-5} = e^{-10}$.*

Proof. For $x < 5$, $X \cdot \mathbb{I}(X \in [5, 10]) > x$ whenever $X \in [5, 10]$ (because this guarantees that $X > x$, and also that the multiplier is one and not zero). Therefore, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = \Pr[X \in [5, 10]]$. Observe also that $\Pr[X \in [5, 10]] = \Pr[X \geq 5] - \Pr[X > 10] = 1 - F(5) - (1 - F(10)) = F(10) - F(5) = e^{-5} - e^{-10}$. ■

Observation 2 *For all* $x > 10$, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = 0$.

Proof. Whenever $X > 10$, the multiplier $\mathbb{I}(X \in [5, 10]) = 0$, so the whole term can never be > 10 (because whenever $X > 10$, it is zeroed out anyway). ■

Observation 3 *For* $x \in [5, 10]$, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = e^{-x} - e^{-10}$.

Proof. To see this, observe that when $x \in [5, 10]$, $X \cdot \mathbb{I}(X \in [5, 10]) > x$ if and only if $X \in [x, 10]$. This is because we first need $X > x$, but we also need $\mathbb{I}(X \in [5, 10]) = 1$. Therefore, we get that $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = \Pr[X \in [x, 10]]$. We again compute this as $\Pr[X \geq x] - \Pr[X \geq 10] = e^{-x} - e^{-10}$. ■

Now that we know $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x]$ for all x, we can compute its expectation:

$$\begin{aligned}
\mathbb{E}[X \cdot \mathbb{I}(X \in [5, 10])] &= \int_0^\infty \Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] dx \\
&= \int_0^5 e^{-5} - e^{-10} dx + \int_5^{10} e^{-x} - e^{-10} dx + \int_{10}^\infty 0 dx \\
&= 5 \cdot (e^{-5} - e^{-10}) + -e^{-x} - xe^{-10}|_5^1 0 \\
&= 5 \cdot (e^{-5} - e^{-10}) + e^{-5} - e^{-10} - 5e^{-10}|_5^1 0 \\
&= 6e^{-5} - 11e^{-10}.
\end{aligned}$$

Finally, to compute $\mathbb{E}[X | X \in [5, 10]] = \mathbb{E}[X \cdot \mathbb{I}(X \in [5, 10])] / \Pr[X \in [5, 10]]$ we get:

$$\frac{6e^{-5} - 11e^{-10}}{e^{-5} - e^{-10}} = 6 - 5 \cdot \frac{1}{e^5 - 1}.$$

■

1.7 Coupling Arguments

Coupling arguments are typically useful to relate two probabilities. I won't give a formal definition here, but I'll try to set up structure for how to use one (but this section doesn't offer tips for when a coupling argument might be appropriate). It will be easiest to start with two examples.

Example 1: Biased Coins.[2] Consider two biased coins. Coin A lands heads with probability p, and coin B lands heads with probability $q > p$. Prove that, for all k, n, the probability that n independent flips from coin B yields $\geq k$ heads is at least the probability that n independent flips from coin A yields $\geq k$ heads.

At first glance, the claim probably seems obvious: each flip of B is more likely to yield heads, so n flips of B should be more likely to yield more heads. But this is not a rigorous proof. One option for a rigorous proof is to explicitly compute the probability of seeing $\geq k$ heads from each coin, and prove that the term is larger for B (possibly by taking a derivative). There is nothing wrong with this approach, but it will wind up with a lot of calculations for what feels like a very obvious claim. A coupling argument, on the other hand, will more closely match the obvious intution. Let's consider the following way to *couple* coin flips for A and B.

[2] I found this example on the Wikipedia page for Coupling arguments. The page is helpful, but perhaps notation-heavy.

- Flip coin A n times, so that each is heads with probability p independently, and call the results $a_1, \ldots, a_n$.
- Use the labels $b_1, \ldots, b_n$ to denote the results of the n flips of coin B.
- For each i:
  - If a_i is heads, set b_i to heads.
  - If a_i is tails, set b_i to heads with probability $(q-p)/(1-p)$, and to tails with probability $(1-q)/(1-p)$.

From here we want to show two things. First, we want to show that the above procedure is a valid way of flipping coin A n times (this is trivial: we explicitly defined each coin to be heads with the correct probability, independently), and also a valid way of flipping coin B n times (this is less trivial, but not so hard). Second, we want to show that when the flips are coupled this way, there are $\geq k$ heads among the A flips only if there are $\geq k$ heads among the B flips (this is fairly straight-forward, but we'll write it out).

First, it's indeed easy to see that this is a valid way of flipping coin A n times, because we explicitly flip coin A n times and each is heads with probability p. Next, observe that coin b_i is heads whenever a_i is heads (which occurs with probability p), or $(q-p)/(1-p)$ of the time when a_i is tails (which occurs with probability $1-p$). Therefore, coin b_i is heads with probability $p+(1-p)\cdot(q-p)/(1-p) = p+q-p = q$. Therefore, we have indeed flipped n coins independently that are heads with probability q, and this is a valid way to flip coin B n times. Finally, observe that there are clearly more heads among the B coins than the A coins always, as whenever coin a_i is heads, so is b_i. These are the main steps needed to complete the proof.

To wrap up the proof, simply observe that the probability of having $\geq k$ heads on n flips from A is exactly the probability of having $\geq k$ heads on n flips among $a_1, \ldots, a_n$ in the process above. The probability of having $\geq k$ heads on n flips from B is exactly the probability of having $\geq k$ on n flips among $b_1, \ldots, b_n$ in the process above. By the work in the previous paragraph, the latter is at least as large as the former, and this completes the proof.

Example 2: Random Walks. Consider an ant that starts at time $t = 0$ at location 0. Every timestep, they either move up $(+1)$ with probability $1/2$ or down (-1) with probability $1/2$, and they stop after n steps, where n is even. Let H_k denote the probability that the ant reaches $+k$ at *some point* during the walk, and let L_k denote the probability that the ant *finishes* their walk at $+k$ or higher. Prove that $H_k = 2L_k$ for all odd k.

For this problem, even the "raw calculations" approach is quite tricky (because it's cumbersome to write a closed form for H_k). However, there is an elegant coupling argument that makes the proof clean. Consider the following process for producing two random walks:

- Let $R_0 = S_0 = 0$. For all i from 1 to n, let R_i be equal to $R_{i-1} + 1$ with probability $1/2$, and equal to $R_{i-1} - 1$ with probability $1/2$ (independently).
- If there exists an i such that $R_i = k$, let i^* be the largest such i. If there is no such i, let $S_i := R_i$ for all i. Observe that because n is even and k is odd, that we must have $k < n$.
- Otherwise, let $S_i := R_i$ for all $i \leq i^*$, and $S_i := k - (R_i - k)$ for all $i > i^*$. That is, "reflect" R_i over the horizontal line at k to get S_i when $i > i^*$.

Again, we need to show two things. First, we need to show that both R and S are correctly sampled random walks. Again, this will be trivial for R, and not-bad-but-not-trivial for S. Then, we need to use these two walks to reason about H_k and L_k.

First, it's clear that R is a correctly-drawn random walk, because we explicitly define it as such. It's also true that S is a correctly-drawn random walk. To see this, observe importantly that *reflecting from R to S after i^* does not change i^**. Indeed, after i^*, R is either completely above k, or completely below k (by definition). Reflecting R to S maintains $S_{i^*} = k$, and also maintains that S is completely below k or completely above k (the opposite of R). This means that the mapping from R to S is its own inverse (applying it twice returns the original R). In particular, this means that every possible random walk is equally likely to be drawn from R (if it is drawn directly) as it is drawn from S (if it's reflection is drawn from R). Therefore, S is also a correctly-drawn random walk.

Finally, observe that either: *neither* R nor S ever reach k, or *both* R and S reach k. Additionally, when neither R nor S reach k, neither R nor S finish above k. On the other hand, when both R and S reach k *exactly one of them finishes above* k. Because both R and S are correctly-drawn random walks, this means that H_k is exactly half the probability that R reaches k, plus half the probability that S reachest k. Similarly, L_k is exactly half the probability that R finishes above k, plus half the probability that S finishes above k. By the reasoning at the beginning of this paragraph, H_k is now exactly twice L_k.

Concluding Thoughts: Coupling arguments are a nice tool to make elegant arguments that better align with intuition than "raw calculations." You won't have to use them extensively in 445, but they will be relevant when we discuss information cascades, and it's a good tool to have in your toolkit to avoid messy calculations.

1.8 "Principle of Deferred Decisions"

The "principle of deferred decisions" isn't a formal theorem or definition, but a concept that will be useful throughout this class to simplify analyses. Consider the following example: say that there are n people who have purchased tickets for a plane. The plane has n seats, and each person purchases a ticket for a distinct seat. The passengers line up in order from 1 to n to board, but unfortunately the first passenger insists on taking a seat that is not their own, uniformly at random. Fortunately the remaining passengers are flexible: if their assigned seat is taken, they will happily take a uniformly random available seat (if their assigned seat is available, they will take it). We wish to understand: *what is the probability that the n^{th} passenger sits in their own seat*?

At first glance, this seems like it should require some messy math. There's a lot of interaction going on between the different passengers (e.g. we first need to know which seat the first passenger takes in order to know which future passengers' decisions matter). But fortunately there's a (conceptually more challenging, but mathematically) simpler way to reason about this. To really appreciate the simplicity which will come later, try to analyze this probability without the tips below (we will not do this exercise in these notes).

Let's first make the following observation (which is not related to the principle of deferred decisions, but to this specific problem): the last passenger will sit either in their own seat, or in seat one. This is because no other seat can be empty: for all $i \in [2, n-1]$, if seat i is available when passenger i boards, they will take it (and fill it). Otherwise, it is already filled. Clearly, exactly one of these two seats will be available when passenger n boards, and we just need to find out which one it is.

Now we can use the principle of deferred decisions. Every time a new passenger i boards the

plane, their seat is either taken or not taken. If their seat is not taken, we ignore them. If seat n is already taken, we ignore them (because certainly passenger n will not sit in their own seat). If 1 is already taken, we ignore them (because now certainly passenger n will sit in their own seat). If their seat is taken and $1, n$ are not, then there are $n - i + 1$ remaining seats, and we need to pick one uniformly at random. One way to pick a uniformly random element from a set of size $n-i+1$ is to just pick an element directly, each with probability $1/(n - i + 1)$. Another way (which may initially seem odd) is to first pick whether the element is in the set $\{1, n\}$ or not. The element is in the set $\{1, n\}$ with probability $2/(n - i + 1)$, and not otherwise. If the element is in the set $\{1, n\}$, then we are about to decide what happens with the last passenger! Conditioned on being in $\{1, n\}$, the probability of selecting 1 is $1/2$ (and n is also $1/2$). So we see that **we are equally likely to select 1 or n, and therefore we are equally likely to have the last passenger sit in their own seat or 1.** If the element is not in the set $\{1, n\}$, we pick the seat uniformly at random from the remainder, and continue. Importantly, **at whatever step we finalize where passenger n will sit, it is equally likely that passenger n gets their own seat or 1.** Therefore, the probability that they get their own seat is exactly $1/2$.

Let's try to unpack the main idea used. One way to draw a uniformly random element of $\{1, \ldots, n\}$ is to just draw each element with probability $1/n$. Another option is to first draw whether the element is even or odd (say n is even). The element is even with probability $1/2$, and odd with probability $1/2$. Then, if we decide that the element is even, we could draw a uniformly random even number. This process will indeed select every even number with probability $1/n$, and also every odd number with probability $1/n$, as desired. The principle gets its name because we are *deferring* the decision of exactly what element to draw until after we have first decided whether it is even or odd. In the abstract, this probably seems not particularly useful, but it gave us a (conceptually challenging, but mathematically) simple proof above. We'll also revisit this in the "basic proof writing" section.

2 Basic continuous optimization

2.1 Single-variable, unconstrained optimization

Say we want to find the global maximum of a continuous, differentiable function $f(\cdot)$. Any value that is a global maximum must also be a critical point, point where $f'(x) = 0$. Not all critical points are local optima, and not all local optima are local maxima, but all local maxima are critical points. One also needs to confirm that $f(\cdot)$ indeed achieves its global maximum by examining $\lim_{x\to\pm\infty} f(x)$.

Example 1: For example, say we want to find the global maximum of $f(x) = x^2$. There is a unique critical point at $x = 0$. So if the function attains its global maximum, it must be at $x = 0$. However, $\lim_{x\to\infty} x^2 = \infty$, so the function doesn't attain its global maximum.

Example 2: Say we want to find the global maximum of $f(x) = 4x - x^2$. The derivative is $4 - 2x$, so there is a unique critical point at $x = 2$. So if there is a global maximum, it must be $x = 2$. We can verify that $\lim_{x\to\pm\infty} = -\infty$, so $x = 2$ must be the global maximum.[3]

[3]We can also verify that $x = 2$ is a local maximum by computing $f''(2) = -2$, but this isn't necessary.

2.2 Single-variable, constrained optimization

Say now we want to find the constrained maximum of a differentiable function $f(\cdot)$ over the interval $[a, b]$. Now, any value that is the constrained maximum must either be a critical point, or an endpoint of the interval. Here are a few approaches to find the constrained maximum:

- Find all critical points, compute $f(a), f(b), f(x)$ for all critical points x and output the largest.
- Confirm that $f'(a) > 0$ (that is, f is increasing at a) and $f'(b) < 0$. This proves that neither a nor b can be the global maximum. Then compute $f(x)$ for all critical points x and output the largest.
- In either of the above, rather than directly comparing $f(x)$ to $f(y)$, one can instead prove that $f'(z) \geq 0$ on the entire interval $[x, y]$ to conclude that $f(y) \geq f(x)$.
- Prove that x is a global *unconstrained* maximum of $f(\cdot)$, and observe that $x \in [a, b]$.

There are many other approaches. The point is that at the end of the day, you must directly or indirectly compare all critical points and all endpoints. You don't have to directly compute $f(\cdot)$ at all of these values (the bullets above provide some shortcuts), but you must at least indirectly compare them. For this class, it is OK to just describe your approach without writing down the entire calculations (as in the following examples).

Example 3: Say we want to find the constrained maximum of $f(x) = x^2$ on the interval $[3, 8]$. f has no critical points on this range, so the maximum must be either 3 or 8. $f'(x) = 2x > 0$ on this entire interval, so therefore the maximum must be 8.

Example 4: Say we want to find the constrained maximum of $f(x) = 3x^2 - x^3$ on the interval $[-2, 3]$. $f'(x) = 6x - 3x^2$, and therefore f has critical points at 0 and 2. So we need to (at least indirectly) consider $-2, 0, 2, 3$. We see that $f'(x) \leq 0$ on $[-2, 0]$, so we can immediately rule out 0. We also see that $f'(x) \leq 0$ on $[2, 3]$, so we can immediately rule out 3, and we only need to compare -2 and 2. We can also immediately see that $f(-x) > f(x)$ for all $x > 0$, and therefore $f(-2)$ is the global constrained maximum.

Example 5: Say we want to find the constrained maximum of $f(x) = 4x - x^2$ on the interval $[-8, 5]$. We already proved above that $x = 2$ is the global *unconstrained* maximum (Example 2). Therefore $x = 2$ is also the global constrained maximum on $[-8, 5]$.

Warning! An incorrect approach. It might be tempting to try the following approach: First, find all local maxima of $f(\cdot)$. Call this set X. Then, check to see which elements of X lie in $[a, b]$. Call them Y. Then, output the argmax of $f(x)$ over all $x \in Y$. This approach *does not work*, and in fact we already saw a counterexample. Say we want to find the constrained maximum of $f(x) = 3x^2 - x^3$ on the interval $[-2, 3]$. Then $f'(x) = 6x - 3x^2$, and f has critical points at 0 and 2. We can verify that $x = 0$ is a local minimum and $x = 2$ is a local maximum. So $x = 2$ is the unique local maximum, and it also lies in $[-2, 3]$. But, we saw that it's incorrect to conclude that therefore $x = 2$ is the constrained global maximum.

2.3 Multi-variable, unconstrained optimization

Say now we want to find the unconstrained global maximum of a differentiable multi-variate function $f(\cdot, \cdot, \ldots, \cdot)$. Again, any value that is the unconstrained maximum must be a critical point,

where a critical point has $\frac{\partial f(\vec{x})}{\partial x_i} = 0$ for all i. Again, not all critical points are local optima/maxima, but all local maxima are definitely critical points. One also needs to confirm that $f(\cdot)$ indeed achieves its global maximum by examining limits towards ∞. Doing this formally can sometimes be tedious, but in this class we'll only see cases where this is straight-forward.[4] Sometimes, it might also be helpful to think of some variables as being fixed, and solve successive single-variable optimization problems. Here are some examples that you might reasonably need to solve:

Example 6: Say you want to maximize $f(x_1, x_2) = x_1 - x_1^2 - x_2^2$. We can immediately see that for any x_1, $f(x_1, x_2)$ is maximized at $x_2 = 0$ (this is what we mean by thinking of x_1 as fixed and solving a single-variable optimization problem for x_2). Once we've set $x_2 = 0$, we now just want to maximize $x_1 - x_1^2$, which is achieved at $x_1 = 1/2$. So the unconstrained maximizer is $(1/2, 0)$.

Example 7: Say you want to maximize $f(x_1, x_2) = x_1 x_2 - x_1^2 - x_2^2$. We can again think of x_1 as fixed and see that $\frac{\partial f}{\partial x_2} = x_1 - 2x_2$, and so for fixed x_1, the unique maximizer is at $x_2 = x_1/2$. We can then just optimize $x_1(x_1/2) - x_1^2 - (x_1/2)^2 = (-3/4) \cdot x_1^2$, which is clearly maximized at $x_1 = 0$. So the unique global maximizer is $(0, 0)$.

Example 8: Say you want to maximize $f(\vec{x}) = \sum_i f_i(x_i)$. That is, the function you're trying to maximize is just the sum of single-variable functions (one for each coordinate of $\vec{x}$). Then we can simply maximize each $f_i(\cdot)$ separately, and let $x_i^* = \arg\max_{x_i}\{f_i(x_i)\}$. Observe that $\vec{x}^*$ must be the maximizer of $f(\vec{x})$. Many instances you will need to solve in the PSets will be of this format.

2.4 Multi-variable, constrained optimization

Finally, say we want to find the constrained global maximum of a differentiable multi-variate function $f(\cdot, \ldots, \cdot)$. Then the same rules as before apply: we must (at least indirectly) consider all critical points and all extreme points. Multi-variable constrained optimization in general is tricky, and would require an entire class to learn enough tricks to solve every instance. Most (possibly all) of the instances you will need to solve in the PSet will be solvable by finding an unconstrained maximizer of f, $\vec{x}^*$, and observing that $\vec{x}^*$ satisfies the constraints.

Example 9: For example, say you want to maximize $f(\vec{x}) = \sum_i x_i e^{-x_i}$, subject to the constraints $-5 \leq x_i \leq 5$ for all i. We can find the unconstrained maximizer by observing that $\frac{\partial f}{\partial x_i} = e^{-x_i} - x_i e^{-x_i}$, which is positive when $x_i < 1$, and negative when $x_i > 1$. So the unique maximizer is at $x_i = 1$. So $(1, \ldots, 1)$ is the unique global maximizer. We observe that $-5 \leq 1 \leq 5$, so $(1, \ldots, 1)$ also satisfies the constraints. So $(1, \ldots, 1)$ is also the constrained maximizer.

Repeat Warning! Again, recall that it is **not** a valid approach to first find all critical points of $f(\cdot)$, and then see which critical points satisfy the constraints and only consider those (recall example at the end of Section 2.2).

3 Basic Problem Solving

The PSets for this class are "short", in that they're only three problems (four if you count the Strategy Designs). But some of the problems will be a full paragraph description, introduce new definitions, etc. **Part of the challenge is figuring out on your own how to break these problems**

[4] Sometimes you'll need to be clever, but ideally very few (if any) proofs will require very tedious calculations.

down into tractable subparts. Problem solving is more of an art, so I can't recommend a concrete step-by-step procedure. However, I can try to give general guidelines/tips. I will use the following problem as a running example for this section:

> Recall that a bipartite graph has two sets of nodes, L and R, with all edges having one endpoint in L and the other in R. Recall also that a perfect matching is a set of edges such that every node is in exactly one edge.
>
> Let G be a bipartite graph with n nodes on each side. Prove that if every node has degree $\geq n/2$, then G has a perfect matching.
>
> **Hint:** You may use Hall's Marriage Theorem. Recall that Hall's Marriage Theorem asserts that a bipartite graph has a perfect matching if and only if for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$, where $N(S)$ denotes the set of nodes with an edge to some node in S.

Step One: Understand what the question is asking. Just because you've taken 240 doesn't mean you're expected to remember off the top of your head exactly what a bipartite graph or perfect matching is, or Hall's Marriage Theorem. The first step is just understanding what all the words mean, and what the question is asking, **and this is part of the assignment**. Normally, all the terms will either be defined in the problem itself, or used in lectures. If that is not the case, you are free to use the Internet to look up terms, and can ask on Piazza if the Internet fails.

In this example, the second paragraph is the actual question. When I read the first sentence, I might get stuck because I don't know what a bipartite graph is. So I should revisit the definition given above. If that's still not clear, I might Google "bipartite graph" to find the Wikipedia page or see some examples. I might even draw some small examples. But I definitely need to know what a bipartite graph is before I can move on to the second sentence, and I should take as much time as necessary to understand the definition. For the second sentence, I may also get stuck on what a "perfect matching" is. I should also refer to the definition given above. Again, if it's not immediately clear, I might draw some small graphs, and look at sets of edges and check whether they satisfy the definition or not. I might also try Google again to see if I can find some examples online.

There is also this hint, which references "Hall's Marriage Theorem." Probably I should know what that means before I start thinking. The statement itself might take some time to parse, because it introduces new notation $N(S)$. But after a few slow reads, I can figure out that every set $S \subseteq L$ must have at least as many neighbors in R as there are elements in S. I might also realize that this makes sense as a necessary condition: if $|N(S)| < |S|$ for any S, then clearly I can't match every node in S to a distinct neighbor, because there are only $< |S|$ possibilities. What's surprising is that this is a sufficient condition, but Hall proved that so we don't need to worry about it.

Some hints are useful to get your thinking started. Other hints are useful to help you formalize the proof once you get the right idea. It's your job to figure out how to use the hint (if you choose to use it at all). In this case, it feels like I probably want to use the hint to get started, since it gives an equivalent condition to G having a perfect matching.

In general, the main thing I want to emphasize here is that you should expect it to sometimes take several minutes just to understand what the question is asking, and that figuring out what the question is asking is part of the challenge. You may need to reread each sentence several times to really understand what it's saying before moving on, and you shouldn't expect to figure it out on your first read-through. It's of course possible to get stuck on this step, and it's fine to use office hours/Piazza for help getting unstuck.

Step Two: Try Stuff! Once you understand what the question is asking, there's no recipe for what to do next. So you have to just play around until inspiration hits. General tips are: try to construct a counterexample and see where you get stuck, or working through small examples (or both). Let's try a small example first, with $n = 2$. **But it will generally not be obvious what you should try.**

When $n = 2$, I can't help but notice that Hall's Theorem seems especially easy to use, since there are only two nodes on the Left (call them u and v, and let the Right nodes be $\{x, w\}$), and therefore only three non-empty subsets to consider. Moreover, we are guaranteed that u has at least one neighbor, so Hall's condition is satisfied. We are also guaranteed that v has at least one neighbor. But we aren't immediately guaranteed that $\{u, v\}$ has two neighbors: What if both have an edge to the same w on Right? But now we see that, if this were the case, then x would have zero neighbors, violating the hypothesis that every node has at least $n/2$ neighbors. Therefore, both x, w must be neighbors of $\{u, v\}$, and Hall's conditions are satisfied for all sets, so there must be a perfect matching.

This gives me some intuition for how to go about the general case. First, it leads me to guess that I should indeed try to use Hall's Theorem right away. Second, it suggests that I need to try and reason about different sets using different logic, because we reasoned about u and v differently than $\{u, v\}$. So let's jump to the general case and see what happens.

Consider an arbitrary n, and we want to show that when every node has degree $\geq n/2$, that every set S on the left has at least $|S|$ neighbors. I also can't help but notice, similarly to the $n = 2$ case, that some sets S will be easy to reason about: if $|S| \leq n/2$, then *just a single node in* S already has $\geq n/2$ neighbors, so clearly S has $\geq n/2$ neighbors and Hall's condition is satisfied. This seems exciting, so let's try thinking about $|S| > n/2$ now. This feels trickier: I do know that there must be $\geq n/2$ neighbors of S, but what if every node in S has the same neighbors? Let's consider what must happen for this to occur. Let $N(S)$ denote the neighbors of S, and let T denote the remaining nodes in Right. We have at our disposal that $|S| > n/2$, and that $|N(S)| < |S|$, and therefore T is non-empty.

From here, I might again get quickly stuck, since the next step isn't obvious. Maybe I would try some random things, or maybe I would revisit the $n = 2$ case, or maybe I'd try to draw out an example with $n = 3$ or $n = 4$ with a particular S with $|S| > n/2$. But eventually I'd get to the realization that any node in T has *no neighbors in* S, which accounts for $> n/2$ of the nodes in Left. Therefore, any node in T has $< n/2$ neighbors, which would contradict our hypothesis. Therefore, T must be empty, which in turn means that we cannot have $|N(S)| < |S|$ for any S of size $> n/2$, which covers all remaining cases.

The two main ideas I want to convey with this example are: (a) you're not supposed to know exactly what the question is asking on your first read-through. It might take several minutes just to figure that out, and (b) you're not supposed to know exactly what step to take next to solve the problem, even if you understand the lectures perfectly. You're supposed to get stuck, and there's no recipe for getting unstuck. I tried to give a few tips above (work through small examples, try to handle special cases first), but it's mostly an art that you learn by doing (the PSets).

Also, please note that the writing above is me trying to explain my thought process. This is not how I would write up a proof once I figure it out (in particular, I would skip the $n = 2$ case, write clearer statements, etc.). See Section 4 for how I would write up a solution to this problem!

3.1 Partial Progress and Partial Credit

Assignments are hard, and you're not expected to fully solve every problem. Section **??** gives some tips on approaches that will get you partial progress. Here, I just want to elaborate on "productive" partial progress versus "unproductive" partial progress.

Here is a thought experiment: if you were to hand your solution to another student, who had not yet solved the problem, how much closer to solving the problem would they be after reading your solution? If the student would not be much further along, I would say you have not made much partial progress (even if you tried very hard, and even if you wrote a lot). If the student would be much further along, then I would say you've made much partial progress. Consider the following two partial solutions to the above problem.

> I was unable to solve the general case, but here is a proof for $n = 2$. Let u, v denote the two nodes on the left. Observe that Hall's Theorem says that a perfect matching exists if and only if: (a) u has at least one neighbor, (b) v has at least one neighbor, and (c) the set $\{u, v\}$ has two neighbors.
>
> Observe that, immediately because every node has degree at least $n/2$, that both u and v must have at least one neighbor, so this covers (a) and (b). To consider $\{u, v\}$, first observe that both right nodes have at least one neighbor (again, immediately from the problem statement). Because $\{u, v\}$ are the only possible neighbors this means that each right node has a neighbor in $\{u, v\}$. In particular, this means that the set $\{u, v\}$ has both right nodes as neighbors, which covers (c).
>
> In the general case, we can again seek to apply Hall's theorem. For any set S with $|S| \leq n/2$, observe that even a single node in S has $n/2$ neighbors, which implies that $|N(S)| \geq n/2 \geq |S|$, as desired. But I can't figure out how to handle sets with $|S| > n/2$.

This solution makes a lot of concrete partial progress. First, it provides a complete proof in the case of $n = 2$. Second, it provides the easy half of a proof for the general case. If I were to grade this solution, I might give it a 14/20.

> Below is a picture of a graph on 4 nodes without a perfect matching (imagine that a picture is given). We note that it has no perfect matching because it violates Hall's Theorem. We also note node u has < 2 neighbors, violating the problem hypothesis. It seems like it would be really challenging to add a neighbor to u without creating a perfect matching. For example, if we add any single edge to u, this yields a perfect matching. Of course, maybe we can add an edge to u and delete another edge, but this also seems unlikely to work.
>
> Also recall that Hall's Theorem can be proved using max-flow-min-cut. As such, we could also consider a proof approach by writing out a network with a source on the left with capacity 1 edge to all left nodes, infinite weight edges from left nodes to right nodes, and capacity 1 edges to a sink on the right. We could then try to prove that whenever each edge has degree at least $n/2$ that there is a flow of weight n, and therefore a perfect matching.

This solution conveys that the author clearly has a lot of ideas. I think it is a good idea to write down a concrete example and play around with it. I also think it's a good idea to see whether max-flow-min-cut helps at all with a proof. But unfortunately there's not much else here. While I'm convinced that the author tried, and did come up with some ideas, none of the ideas make concrete progress towards a solution. Put another way, the solution does convince me that the author is further along towards a solution than when they first started, but the solution wouldn't help another problem-solver. If I were to grade this solution, I might give it a 6/20.

The main point I want to make in this section is the thought experiment: you should measure your partial progress by "how much would this writeup help another student?" and not by "how

hard does this writeup prove that I've worked?" I hope that the examples above help display the difference.[5]

4 Basic Proof Writing

This is a bit of an oversimplification, but I think there are two 'kinds' of proof-writing that this class will develop. First, you must be able to write rigorous, complete proofs of short claims. Second, you must be able to write a clear, rigorous outline for a complex proof, by breaking it down into concrete, rigorous claims. Section 4.1 deals with the first kind, and Section 4.2 deals with the second. **I strongly recommend reading both sections.**

4.1 Writing short proofs

I found the following source: `https://math.dartmouth.edu/archive/m31x12/public_html/Proof%20Writing.pdf` to be a good quick source for tips on writing a proof. In particular, this source notes that while a good proof should be written in complete sentences (and not a sequence of formal mathematical statements), it should still be possible for a reader to understand what is the sequence of formal mathematical statements which corresponds to your complete sentences. I'll also include below a few generals pitfalls I noticed in previous semesters. Note that most proofs you'll write in 445 are much longer than the examples below, but hopefully it is enough to give a sense of what these pitfalls might look like.

Pitfall One: False Implications. The most common reason that a proof is incorrect is that there is a false implication along the way. For example, let's revisit the incorrect approach under single-variate optimization. If I were trying to find the maximum of $f(x) := 3x^2 - x^3$ on the interval $[-2, 3]$ and wrote the following:

> The derivative of $f(x)$ is $f'(x) = 6x - 3x^2$. There are two critical points: $x = 0$ and $x = 2$. As $f''(x) = 6 - 6x$, we see that $x = 0$ is a local minimum, and $x = 2$ is a local maximum. Because $x = 2$ is the only local maximum, it must also be the constrained global maximum.

This proof is "obviously" incorrect, because it claims that $x = 2$ is the constrained global maximum (when it is $x = -2$). Let me change the example slightly, so that we are trying to are trying to find the maximum of $f(x) := 3x^2 - x^3$ on the interval $[0, 3]$, and repeat the same proof, word for word:

> The derivative of $f(x)$ is $f'(x) = 6x - 3x^2$. There are two critical points: $x = 0$ and $x = 2$. As $f''(x) = 6 - 6x$, we see that $x = 0$ is a local minimum, and $x = 2$ is a local maximum. Because $x = 2$ is the only local maximum, it must also be the constrained global maximum.

Even though $x = 2$ is indeed the constrained global maximum (so the "solution" is correct), the proof is still incorrect, but it's now harder to see why. The very last line of the "proof" is combining two logical statements together. First, it is claiming that every constrained maximum must also be a local maximum. Second, it is observing that because there is only one local maximum, it must be the constrained global maximum (otherwise, the constrained maximum would not be a local

[5] All examples in this document were made up by me for the purpose of explanation — these are not actual problems I've asked in 445, nor actual student solutions.

maximum, contradicting the first claim). But the first of these logical claims is false. Indeed, we just saw an example above where $x = 2$ is the only local maximum, but is *not* the constrained global maximum (and therefore, the constrained global maximum is not a local maximum).

So to summarize the flaw here, when we mapped the last sentence of the proof into the corresponding logical claim, that claim was false, and therefore the proof is incorrect (even though the final conclusion happens to be true). As a general rule: if the same proof, verbatim, could be used to prove a statement that is false, then the proof is certainly incorrect. Similarly, if the same sentence, verbatim, could be inserted into an otherwise correct proof of a false statement, then that sentence is certainly incorrect.

Pitfall Two: Overly Vague Implications. Another reason for proof to be incorrect is that it is not possible for the reader to map the complete sentences to logical claims. In particular, maybe the sentence is too imprecise, and it could reasonably be intending to make many possible logical claims, some of which are false. For the same example (again, proving that $x = 2$ is the constrained global maximum for $f(x) := 3x^2 - x^3$ on $[0, 3]$), consider the following sentence:

> $x = 2$ is a critical point, with negative derivative to the right and positive derivative to the left. Therefore, $x = 2$ is the constrained global maximum.

A favorable interpretation of this sentence is that the author is claiming that because the derivative is negative on $[2, 3]$ (to the right), and positive on $[0, 2]$ (to the left), then $x = 2$ must be the constrained global maximum. This is a correct argument. However, an equally reasonable interpretation of this sentence is that the author is claiming that because the derivative is negative on some interval $[2, z)$ (to the right), and positive on some interval $(y, 2]$ (to the left), then $x = 2$ is a constrained global maximum. This argument is false (as it only proves that that x is a local maximum).

Depending on the surrounding context (e.g. if the author states prior to this sentence that "to the left" means "all the way left until the end of the interval"), maybe this particular sentence could be clear. But in isolation, the underlying logical claim is unclear. A similar general rule applies here: if your English sentence could reasonably be mapped (taking the surrounding context into account) into a logical claim that is false, then it is incorrect.[6]

Pitfall Three: Too Many Missing Steps. In 445, you are *certainly* not expected to "show your work" for mundane calculations. But it is still possible to pack too many logical claims into the same short sentence in a way that the reader has no hope of following. There's no objective measure for what counts as too many, but you should expect to learn throughout the semester (via the lecture notes, staff solutions to PSets, other students' solutions to PSets that you see on MTA) what is considered sufficient detail. In general, my goal is for the staff solutions to provide a little bit more detail than what is necessary for full credit.

For example, for 445, the following would be plenty sufficient to prove that the global maximum of $f(x) := 3x^2 - x^3$ on $[0, 3]$ is $x = 2$.

> $f'(x) = 6x - 3x^2$, and therefore the only critical points are 0 and 2. Note that $f'(x) \leq 0$ on $[0, 2]$, and $f'(x) \geq 0$ on $[2, 3]$. Therefore, $x = 2$ is the constrained global maximum.

Let me give a slightly different example, say that you are given the word problem: Alice sells apples, and knows that if she sets price $x \in [0, 3]$ per apple, then exactly $f(x) := 3x - x^2$ apples will

[6] Of course, the 445 graders will try to read anything you write on the favorable side of reasonable. But it certainly does not mean that sentences with multiple interpretations will always be given the most favorable one.

be purchased. What price x should Alice set to maximize her total revenue from all sold apples? Then the following answer is *not* sufficient, because it skips too many steps:

> The constrained global max of $xf(x)$ on $[0, 3]$ is $x = 2$, so Alice should set price 2.

The above answer forces the reader to do too much in their head. There is nothing factually inaccurate about the above proof, but it essentially skips the first half of the problem (why is a global maximum of $xf(x)$ relevant?). A better solution is below:

> If Alice sets price x per apple, and $f(x)$ apples are purchased, then her revenue for setting price x is $xf(x) := 3x^2 - x^3$. As Alice wishes to maximize her revenue, she should set the price maximizing $xf(x)$ on $[0, 3]$, which is $x = 2$. To see this, observe that the derivative of $xf(x)$ is positive on $[0, 2]$, and negative on $[2, 3]$.

As a general rule, it should be *easy* for a 445 grader to read your proof, understand what you are saying, and whether or not it is correct.

4.1.1 A Final Example

Consider the following problem, which I first heard about here: `https://gilkalai.wordpress.com/2017/09/08/elchanan-mossels-amazing-dice-paradox-answers-to-tyi-30/`. You roll a fair six-sided die until it lands six. What is the expected number of rolls you make (included the one which lands six), conditioned on all rolls being even? Let's view two conflicting "proofs." In both, we'll use without proof the following fact:

Fact 6 *Let D be a distribution such that when random variable X is drawn from D, the probability that $X = x$ is p. Then if we repeatedly sample draws from D independently until we see one which is equal to x, the expected number of draws we make is $1/p$.*

Proof. This is also a good example where the math is simpler if we use "form 2" of the definition of expectation. The probability that we make strictly more than i draws is the probability that all of the first i draws are not equal to x. Because they are drawn independently, and equal to x with probability p, this is just $(1 - p)^i$. So we get that the expected number of draws we make is $\sum_{i=0}^{\infty}(1 - p)^i = 1/p$. ■

Proof 1:

> We know that if we were to roll the die until we hit a six, it would take six rolls in expectation, by Fact 6 (because we have a $1/6$ chance of rolling a six each time). If instead we condition on all rolls being even, now there are only three possibilities instead of six, so the probability of rolling a six each time is $1/3$ instead of $1/6$. So by the same Fact 6, the expected number of rolls until we hit a six, conditioned on all rolls being even, is now three.

Proof 2:

> Consider instead repeatedly rolling a die in the following manner, using the principle of deferred decisions. First, decide if the die will land on two/four, or not on two/four (then decide exactly the roll, uniformly at random among the remaining possibilities). Stop as soon as the die lands not on two/four. Then the probability of terminating any given round is $2/3$, and so by Fact 6, the expected number of rolls is $3/2$. Moreover, observe that we can decide whether or not to stop rolling independently of whether the last roll is a six or odd. Therefore, the expected number of rolls until we hit a six, conditioned on all rolls being even, is $3/2$.

Both proofs seem tempting: the logic in proof 1 is pretty straight-forward to follow. Proof 2 may be extra tempting because it uses a fancy term that was introduced earlier. Proof 2, it turns out, is correct (but you should not typically associate correctness with fancy terms), and Proof 1 is not. The first sentence of Proof 1 is correct. The second sentence of Proof 1 is vague or incorrect. In particular, when the proof says "now there are only three possibilities instead of six," it seems to suggest that conditioning on all rolls being even is the same as independently rolling each die, and enforcing that each draw is even. These are not the same (likely the original motivation for this problem was to point out this misconception).

Indeed, let's consider instead a million-sided die. The point is that we are *extremely* unlikely to have a long run where all rolls are even, so conditioning on all rolls being even makes the length of the runs quite short. In particular, we shouldn't expect to have all even throws followed by a six for a long run at all, and most of the time when this happens, it's because we got a six very quickly. If we repeat the argument in Proof 1, it would imply that the expected number of throws, conditioned on all throws being even, is 500000. Proof 2 instead suggests that the expected number of throws until we hit a roll which is either odd or six is $1000000/500001 \approx 2$. Hopefully that gives some intuition. We can also do the full calculation to confirm:

The probability that we roll exactly i times until hitting a six, and that the first $i-1$ rolls were all even (i.e. two or four), is $(1/3)^{i-1}/6$. So the probability that we roll all evens until hitting a six is:

$$\sum_{i \geq 1} (1/3)^{i-1}/6 = 1/4.$$

Also, the total number of rolls, only counting those from sequences from which we rolled evens until hitting a six is:

$$\sum_{i \geq 1} i(1/3)^{i-1}/6 = 9/24.$$

The conditional expectation then just divides these two to get $\frac{9/24}{1/4} = 9/6 = 3/2$.

4.2 Effectively breaking down long proofs

Let's revisit the problem from Section 3, and see how to write a clear proof of a complex claim. Recall first the problem:

> Recall that a bipartite graph has two sets of nodes, L and R, with all edges having one endpoint in L and the other in R. Recall also that a perfect matching is a set of edges such that every node is in exactly one edge.
>
> Let G be a bipartite graph with n nodes on each side. Prove that if every node has degree $\geq n/2$, then G has a perfect matching.
>
> **Hint:** You may use Hall's Marriage Theorem. Recall that Hall's Marriage Theorem asserts that a bipartite graph has a perfect matching if and only if for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$, where $N(S)$ denotes the set of nodes with an edge to some node in S.

Here is a solution I would write. Afterwards, I'll explain what I think of as the key points.

Solution. We will prove that G has a perfect matching using Hall's Marriage Theorem, and show that for all sets $S \subseteq L$, $|N(S)| \geq |S|$.

Let us first consider sets S where $|S| \leq n/2$.

Lemma 4 *Consider any $S \subseteq L$, with $|S| \leq n/2$. Then, $|N(S)| \geq |S|$.*

Proof. Let v be any node in S. Observe that, immediately by the definition of G, the degree of v is at least $n/2$. This immediately implies that v has at least $n/2$ neighbors, and therefore S has at least $n/2$ neighbors. Because $|S| \leq n/2$, we have that $|N(S)| \geq n/2 \geq |S|$, as desired. ■

Next, we consider the case where $|S| > n/2$.

Lemma 5 *Consider any $S \subseteq L$, with $|S| > n/2$. Then, $|N(S)| \geq |S|$.*

Proof. In fact, we will prove an even stronger claim, that $|N(S)| = n$. To do this, assume for contradiction that $|N(S)| < n$. This means that there must exist some node $u \in R$ such that $u \notin N(S)$. In particular, this means that u has no neighbors in S. However, there are $> n/2$ nodes in S, and therefore $< n/2$ nodes in $L \setminus S$. This means that the degree of u is strictly less than $n/2$ (because all of u's neighbors must lie in $L \setminus S$). This contradicts the definition of G, as every node has degree at least $n/2$. ■

Now, we can wrap up the proof. Lemmas 4 and 5 together prove that $|N(S)| \geq |S|$ for all $S \subseteq L$. Hall's marriage theorem now implies that G has a perfect matching, as desired.

Thoughts on this writeup. Here is a great test to see if you've successfully broken down your complex proof into clear subparts: Ignore the proofs of Lemma 4 and Lemma 5. Now, the entire proof is just a few sentences. Assuming that Lemma 4, Lemma 5, and Hall's Marriage Theorem are all correct, *is it easy to follow the logical flow?* This is always a bit subjective, but I'd argue that the logic is quite clear, and is entirely captured in the final two sentences. This is the key difference between writing a 'long' proof and a 'short' proof. For a long proof, it's impossible for a reader to follow multiple logical trails at once, so your job is to break it down into manageable short proofs, and also provide a single short proof to bring it all together. Try to think of it like a tree: the root is the outline which connects Lemma 4, Lemma 5 and Hall's Marriage Theorem to prove the claim. It has three children: Lemma 4, Lemma 5, and Hall's Marriage Theorem. Lemma 4 has a standalone proof, because the logic is short and coherent. Lemma 5 has a standalone proof, because the logic is short and coherent. Hall's Marriage Theorem is given to you, and does not require a proof. Each node in the tree should provide a clear, logically coherent proof (i.e. the proof uses just a few ideas, and can fit in the grader's head all at once) of the desired claim, assuming that the claims made in its children are correct. Here are some other bulleted thoughts:

- Is it crucial that the proof separates out the key claims using the Lemma environment in LaTeX? No. But, if you're new to proof-writing, this is a good structure to enforce on yourself. I personally try to use this structure whenever I write my own proofs.
- Is it crucial that Lemmas 4 and 5 are broken down into two lemmas, instead of just one? No. But, the two cases clearly use different logic. So absolutely, they should at least be broken up into separate paragraphs.
- It *is* crucial that each subpart has a clear, concrete, and formal statement. For example, it's crucial that it's easy to see that Lemmas 4 and 5 together cover all S, and connect to Hall's Theorem. Informal statements like "All sets have sufficient neighbors" (what is "sufficient"?) or "small sets satisfy $|N(S)| \geq |S|$" (what is "small"?) fail this, because the grader can't figure out exactly what's proved in these steps.

- Staff solutions and lecture notes will give you further examples of how to break down large proofs into smaller chunks. When you read them, try to go through the exercise of reading just the definitions/lemmas/conclusion, and confirming that the logic follows. Then, when reading each individual lemma, you just need to confirm that this individual lemma is correct.

4.3 And Two Quick Tips

Tip One: I still suggest this even to PhD students, and follow this advice myself: **after you write something up, read it yourself** (perhaps even out-loud-in-your-head). If you can't follow your own logic, don't be surprised when the grader can't either! You will likely be surprised at how much better your own writing can get just by iterating this process.

Tip Two: Remember that it should be *easy* for a 445 grader to follow your solution. Even if your solution is complicated, and it will take a long time to process everything you've written, you should try to write a concise outline, and/or try to state what you're about to do before diving into a pool of calculations.